R E P O R T R E S U M E S

ED 011 045 24

PROBLEM SOLVING PROFICIENCY AMONG ELEMENTARY SCHOOL TEACHERS. II, TEACHERS OF ARITHMETIC, GRADES 3-6.

BY- TURNER, RICHARD L.
INDIANA UNIV., BLOOMINGTON, INST. OF EDUC. RES.

REPORT NUMBER CRP-419-2 PUB DATE JUN 60
EDRS PRICE MF-$0.18 HC-$2.88 72P.

DESCRIPTORS- *ARITHMETIC, *TEACHING, *PROBLEM SOLVING, *TESTING, *TEACHER CHARACTERISTICS, ELEMENTARY SCHOOL TEACHERS, BLOOMINGTON

THE PURPOSE OF THE STUDY WAS TO EXAMINE THE RELIABILITY AND VALIDITY OF SEVEN PROBLEMS IN THE TEACHING OF ARITHMETIC. THE VALIDITY OF THE PROBLEMS WAS INVESTIGATED ON THE CRITERIA OF (1) DIFFERENTIATING ELEMENTARY SCHOOL TEACHERS OF ARITHMETIC, GRADES 3-6, FROM COMPARABLY EDUCATED NONTEACHERS, (2) BEING SENSITIVE TO THE EFFECTS OF ELEMENTARY SCHOOL TEACHING EXPERIENCE, AND (3) HOLDING SENSIBLE RELATIONSHIPS TO NUMEROUS INDEPENDENT VARIABLES. IT WAS FOUND THAT TEACHERS AS A GROUP OUTPERFORM NONTEACHERS AS A GROUP AND THAT PERFORMANCE IS SENSITIVE TO THE EFFECTS OF TEACHING EXPERIENCE. INTELLIGENCE, READING COMPREHENSION, MINNESOTA TEACHERS ATTITUDE INDEX (MTAI) SCORE, AND SIZE OF INSTITUTION AT WHICH A TEACHER PREPARES ARE INDEPENDENT VARIABLES POSITIVELY RELATED TO PERFORMANCE. THE SPLIT-HALF RELIABILITY FOR THE MOST VALID SCORE WAS .84 AMONG TEACHERS AND .87 AMONG STUDENT TEACHERS. THE STABILITY COEFFICIENT FOR STUDENT TEACHERS, WITH A 4-MONTH INTERVAL BETWEEN TESTINGS, WAS .63. FROM THESE RESULTS, THE CONCLUSION WAS DRAWN THAT THE PROBLEMS IN TEACHING ARITHMETIC THAT WERE INVESTIGATED WERE REASONABLY RELIABLE, ALTHOUGH SOMEWHAT WEAK IN STABILITY, AND THAT THEY HELD THE EXPECTED SET OF RELATIONSHIPS TO THE CRITERIA CHOSEN FOR THE INITIAL STEPS IN VALIDATION. (TC)

ED011045

CRP-419 S-1459

MONOGRAPH
of the
INSTITUTE OF EDUCATIONAL RESEARCH
at
INDIANA UNIVERSITY

Problem Solving Proficiency Among Elementary School Teachers.

II. Teachers of Arithmetic, Grades 3-6 .

by
Richard L. Turner

JUNE 1960

PROBLEM SOLVING PROFICIENCY AMONG ELEMENTARY SCHOOL TEACHERS. II, TEACHERS OF ARITHMETIC, GRADES 3-6.

by

Richard L. Turner

The research on which this study was based was supported in part by the Cooperative Research Program of the U. S. Office of Education under Project 419 (7790).

Published by the
INSTITUTE OF EDUCATIONAL RESEARCH
School of Education
Indiana University

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

Price, $1.00, postpaid
For sale by the Institute of Educational Research
School of Education
Bloomington, Indiana

TABLE OF CONTENTS

	Page
INTRODUCTION	1
Purpose	1
Strategy and Rationale	1
The basis of criterion selection	2
The determination of reliabilities	7
PROCEDURES	8
Problem Construction	8
Delimitation of the arithmetic domain	8
Problem development	8
Problem A	11
Problem B	11
Problem E	12
Problem H	12
Problem F	13
Problems D and G	13
Determination of random responses scores	13
Sampling Procedures	15
Teacher sample	15
Non-teacher sample	16

Page

Preparatory teacher sample 16

Method of obtaining data on independent variables . . 18

RESULTS 20

The Differentiation of Teachers from Non-teachers 20

Sample characteristics 20

Validity of criterion scores 22

Sensitivity of the Problems to Differences in Teaching Experience 25

Relationships between Problem Solving Performance and Selected Independent Variables 30

Intelligence 31

MTAI and Study of Values 32

Teaching location 33

Size of Teacher preparatory institution 35

Reading comprehension 37

Grade level 38

Arithmetic methods as a treatment 39

Recency of exposure to arithmetic methods 40

Number of courses in mathematics 42

Sex 44

Error Scores 45

Page

Inter-problem, Inter-score Relationships 47

Reliabilities 47

DISCUSSION 53

SUMMARY AND CONCLUSIONS 59

REFERENCES 60

LIST OF TABLES

Table		Page
1.	Number of Subjects in Each Group	18
2.	Number of Teachers and Non-teachers by Years of Higher Education	20
3.	Number of Teachers and Non-teachers by Number of Courses in Mathematics	21
4.	Number of Teachers and Non-teachers in Six Age Categories	21
5.	Number of Men and Women in the Teacher and Non-teacher Samples	21
6.	Means and Variances of PI Score and t Value: Teachers vs. Non-teachers	22
7.	Means and Variances of VN Score and t Value: Teachers vs. Non-teachers	22
8.	Means and Variances of CR Score and t Value: Teachers vs. Non-teachers	23
9.	Teachers vs. Non-teachers According to Consistency of Performance	24
10.	Proportion of Teachers at Four Levels of Experience Coming from Small, Medium, and Large Preparatory Institutions	26
11.	Analysis of Variance for PI Scores of Teachers with 0, 1-3, 4-10, and 11-25 Years Experience	26
12.	Means and Variances of PI Scores, and t Values for Teachers with 0, 1-3, 4-10, and 11-25 Years Experience	27
13.	Means and Variances of PI Score for Teachers with 0 Years Experience from Large Preparatory Institutions, and Teachers with 4-10 and 11-25 Years Experience from Small Institutions	28
14.	Means and Variances of PI Score, and t Value for Teachers of 1-7 Years Experience, Age 23-29 and 30-49	29

Table		Page
15.	Consistency of Performance, Teachers with 0 Experience vs. Teachers with 1-3 Years Experience	29
16.	Percentage of Teachers and Non-teachers at or above each of Three PI Scores	30
17.	Means and Variances of ACE Raw Scores, and t Value, High Consistency vs. Low Consistency Student Teachers . . .	31
18.	Correlations between the MTAI, the Study of Values, and the PI Score for Two groups of Student Teachers	32
19.	Means and Variances, and t Value of MTAI and Study of Values Scores for Student Teachers of High and Low Consistency.	34
20.	Analysis of Variance of PI Scores for County, Town, and City Teachers	35
21.	Consistency of Performance of Teachers from a Small Consolidated System vs. Teachers from a Large Consolidated System	35
22.	Number of Teachers Prepared at Institutions of Size 0-999, 1000-4999, and 5000 and above at Each of Four PI Score Intervals	36
23.	Number of Teachers Prepared at Institutions of Size 0-999, 1000-4999, and 5000 and above in the High and Low Consistency Groups	36
24.	Means and Variances of Coop C_2 Raw Scores, and t Value High Consistency vs. Low Consistency Student Teachers .	37
25.	Number of Teachers in Four PI Score Intervals When Classified by Grade Levels Taught	39
26.	Means and Variances of PI Score for Matched Preparatory Teachers before and after Arithmetic Methods	40
27.	Number of Teachers with Arithmetic Methods within the Past Nine Years vs. Those with Methods 10 or more Years . .	40

Table		Page
28.	Analysis of Variance of PI Scores: Size of Institution at which Methods were last Completed by Years since Completion .	41
29.	Number of Teachers with 1-2, 3-4, and 5 or more Mathematics Courses in Four PI Score Intervals	43
30.	Analysis of Variance of PI Scores: Number of Courses in Mathematics by Grade Level Taught	43
31.	Number of Each Sex Falling at Four PI Score Intervals	44
32.	Number of Each Sex in High and Low Consistency Groups . . .	44
33.	Patterns of Error Responses among High Consistency and Low Consistency Teachers, and Low Consistency Non-teachers.	45
34.	Inter-problem, Inter-score Correlations for 41 Teachers in the Consolidated Town-County System	48
35.	Inter-problem, Inter-score Correlations for 95 Teachers in the Consolidated City-County System	49
36.	Means and Variances of PI Score, and t Values for Three Groups of Student Teachers, One Group Measured before and One Group Measured before and after Student Teaching	50
37.	Reliability Coefficients by Groups, Types, and Scores	51
38.	Differences in Number of Problems Solved above Chance Level Between First and Second Testing of Student Teachers . .	52

INTRODUCTION

Purpose

The objective of the present study is to examine the reliability and validity of seven paper and pencil type problems in the teaching of arithmetic, grades 3-6.

Strategy and Rationale

The research strategy under which the problems in teaching arithmetic were developed has been set forth elsewhere by Turner and Fattu (14). Centrally, this strategy involves the development of problems and the assessment of problem solving proficiency among elementary school teachers in many areas of instruction, of which arithmetic is only one. The objectives of the strategy are to identify the teachers most proficient at solving teaching problems and to identify the characteristics of these teachers. The role played by this strategy in the present study was primarily to clarify the classes of variables which might be explored in a pilot study. The specific procedural rationale upon which the present study is based appears partly in the discussion of strategy but primarily in a second publication, "Problem Solving Proficiency Among Elementary School Teachers I. The Development of Criteria" (15).

Under the strategy and rationale heretofore set forth, the first step in research involves the construction and validation of teaching problems, since valid teaching problems are a necessary condition of the workability of the rationale and the success of the strategy. Of the several criteria which might have been chosen from the rationale for initial validation of problems in teaching arithmetic, the author chose three: (i) the degree to which the problems

differentiate teachers of arithmetic grades 3-6 from comparably educated non-teachers, (ii) the sensitivity of the problems to the effects of teaching experience, and (iii) the relationships held by performance on the problems to selected independent variables, i.e. intelligence, reading comprehension, attitudes toward children, personal values, number of courses in mathematics, arithmetic methods (as a treatment), years since arithmetic methods were completed, size of graduating institution, location of school in which teaching is done, grade taught, age, and sex.

The basis of criterion selection. The degree to which the problems differentiate arithmetic teachers grades 3-6 from comparably educated non-teachers was selected as a criterion on the basis of two propositions. The first proposition is that a valid test of the skills of a specialized group must be capable of differentiating this group from other persons. Unless the test in question can perform this function it must either be conceded that elementary teachers of arithmetic are not specialized sufficiently to be differentiable from the general college graduate population or that the test does not measure specialized skills. Since the strategy under which the author is operating implicitly assumes that teachers are specialized and since the problems are held to measure specialized skills, a failure to obtain the required differentiation would reduce confidence not only in the validity of the problems used but also in the strategy of which they are a part.

A second proposition bearing on the differentiation of elementary arithmetic teachers from comparably educated non-teachers of arithmetic is

that this criterion is independent of all hypotheses and assumptions in the rationale which undergirds the present study, and the hypotheses and assumptions of other, similar, rationales except the assumption that teachers are specialized. Test rationales may differ in how they conceptualize teacher specialization; but no matter how they conceptualize it, each of them must be capable of demonstrating that teachers are in fact specialized, i.e. discriminable from other groups with respect to the skills they possess. In the present study this criterion is the only one that is independent of both the particular rationale adopted and the procedures used in problem construction.

As a criterion, sensitivity of the problems to the effects of teaching experience is not independent of the rationale upon which the problems are based. It was hypothesized in the rationale that other things being equal, the greater the opportunity to acquire instrumental responses relevant to bringing about desirable behavior among pupils, the greater the problem solving proficiency of the teacher will be, up to some asymptote. Classroom experience in teaching arithmetic clearly entails the opportunity to acquire instrumental responses in bringing about desirable arithmetic outcomes with pupils. Constructed problems in teaching arithmetic, then, must be sensitive to the effects of teaching experience. A failure to meet this criterion would imply either that the rationale contains an untenable hypothesis or that the problems are not valid or both. Whether it is the rationale or the problems or both that are at fault cannot be determined from failure to meet this criterion alone.

In addition to meeting the two criteria discussed above a valid test of problems in teaching arithmetic would be expected to hold sensible relationships

to a number of independent variables. Of these variables, intelligence, attitudes toward children, personal values, location of school in which teaching is done, and size of graduating institution were suggested as possibly relevant in the rationale utilized.

It has been previously stated that if other things are equal, opportunity to acquire instrumental responses relevant to bringing about desirable behavioral outcomes with pupils will increase problem solving performance. The acquisition of instrumental responses, however, is obviously not contingent solely on the opportunity to acquire them; it may be hypothesized also to be contingent on learning ability, which may for convenience be equated with intelligence. Given equal opportunity to acquire responses relevant to solving teaching problems, the more intelligent person should acquire a greater number of them.

While intelligence would be expected to be a significant variable in the number of relevant responses acquired given equal opportunity to acquire them, it would also be expected that differences in the perception of what educational outcomes are important would be a significant variable. A teacher would be expected to learn many responses toward those goals which he thinks valuable and characteristically works toward, and very few toward goals which he thinks are unimportant. At present there is no instrument available for making a direct assessment of those particular educational goals a teacher thinks important. There are available, on the other hand, instruments, notably the MTAI (8) and the Allport-Vernon-Lindsey Study of Values (2), which get at the types of pupil behavior a teacher accepts and rejects and at the teachers'

general value orientation. While both of these instruments are peripheral in assessing the arithmetic outcomes a teacher thinks important, they nonetheless provide a means of opening up for exploration the area of teacher values in relation to problem solving performance.

A third variable bearing on the opportunity to acquire instrumental responses is indexed by location in which teaching occurs. The freedom a teacher has in seeking the goals he believes to be important and in utilizing a variety of instructional methods cannot be assumed to be equal in all school systems. The opportunity to acquire particular responses toward arithmetic goals may thus vary according to where the teacher teaches. The direct assessment of the autonomy a teacher has, or the opportunities he has to learn, would require devices for rating or otherwise obtaining indices of the autonomy permitted to a teacher. Such devices are not available; hence, all that can be done at present is to examine differences between teacher performance by school systems as a means of obtaining leads to systems which differ in autonomy permitted and which might serve as a basis for constructing rating scales of autonomy or obtaining other indices of this variable.

The examination of size of teacher preparatory institution as a possibly relevant variable is based in part on the hypothesis that the opportunity to acquire instrumental responses during professional preparation varies between institutions, and in part on the hypothesis that preparatory institutions vary in the educational goals they stress as important and which are learned as important by students. Neither the opportunity to acquire responses in preparation nor the goals stressed in preparatory institutions can be independently assessed

at the present time. This fact again makes it necessary simply to open this area to exploration by examining a crude variable, namely differences between graduates according to institution size.

There are two independent variables whose conceptual relevance hinges on the procedures used in problem construction. First, because paper and pencil problems were constructed, the role of reading comprehension in problem performance must be observed. Clearly if a very high positive relationship were to prevail, there would be grounds for asserting that what purports to be a test of skill in solving teaching problems is in fact a reading test. Second, because the problems do not equally represent each of grades 3, 4, 5, and 6, it is necessary to discover whether performances vary by grade level. If grade 3 teachers, for instance, were to consistently score better than grade 6 teachers, the problems could not be asserted to be equally relevant for all grades in the range suggested.

There are three variables which are conceptually relevant to the arithmetic content of the problems: arithmetic methods as a treatment, number of courses in mathematics, and years since arithmetic methods were last taken. Two of these variables, arithmetic methods as a treatment and number of courses in mathematics, deal only with whether <u>exposure</u> to formal courses in mathematics and arithmetic methods is relevant to problem solving performance. The third variable involves, of course, the <u>recency</u> of exposure to arithmetic methods. These three variables were chosen primarily because they are easily assessable, and, if they are significant, make readily available predictors of problem solving performance. The relationship between more refined

measures of arithmetic knowledge and problem solving performance is dealt with in studies which will be reported at a later time.

In addition to the conceptually relevant variables mentioned above, two other variables, age and sex, were of interest from the viewpoint of identifying sources of variation which might need to be controlled in subsequent studies.

The determination of reliabilities. Since the possible uses of a measure of problem solving proficiency include both prediction and assessment of change after intervening treatments, the stability, i.e. the between occasions reliability, of such a measure is important. Moreover the use of the problems as predictors is most likely to occur with undergraduate preparatory teachers. This population was thus chosen as the one from which to obtain a stability coefficient.

In addition to information on the between occasions reliability of the problems, information on the within-an-occasion reliability is relevant since the error variance within occasions is compounded in the stability coefficient.

PROCEDURES

Problem Construction

Delimitation of the arithmetic domain. The domain of arithmetic objectives in relation to which problems in teaching arithmetic were constructed was limited to those objectives stated by Brownell and published by the National Council of Teachers of Mathematics (4). The arithmetic textbooks of Ginn and Co. (6); Scott, Foresman and Co. (11); World Book Co. (7); John C. Winston and Co. (3); and A Chart for grades 3-8 of the new arithmetic series, Arithmetic We Need (5) were used to determine the appropriateness of given problems for teachers at particular grade levels between grades 3 and 6.

Problem development. In accord with the rationale (14, p. 24), only problems which could easily be administered, and scored on a product criterion (as opposed to process criteria) were constructed for the present study. Over a three year period 20 such problems were constructed and tried out with a cumulative total of about 400 students enrolled in undergraduate and graduate courses in arithmetic instruction. Most of these problems required free, written responses (essay). Problems which were found to be extremely difficult, or which could not be reliably scored, or which yielded small variances for long performance times were eliminated. The remaining problems were utilized to develop the instrument used in the present study. The central concern in the final stages of development was to convert as many as possible of the free, written response problems into problems which could be objectively

scored.

To accomplish this conversion, the written responses for each problem were placed in three classes according to their relevance to the solution of that problem, i.e. relevant, moderately relevant, and not relevant. Responses were then drawn from each category on two criteria, (i) that the response represent a set of the varied responses within a category and (ii) that the degree of relevance of the response was unambiguous with respect to the information available in the problem for making inferences. This procedure yielded a set of alternative responses for each of four problems that had originally been of the written response type. The task for the subject responding to the problems was to decide whether each alternative to each problem stood in the relationship of relevant (A), moderately relevant (B), or not relevant (C), to the solution of the problem. The subject's response was recorded by marking what he believed to be the best response to each alternative.

The scoring of this method of response was dealt with by analogy to the concepts expressed by Meehl and Rosen (12) in a discussion of antecedent probability in relation to the efficiency of cutting scores. A's called A (AA), and B's called B (BB) were designated positive hits or positive identifications (PI) while C's called C (CC) were designated valid eliminations (of non-relevant responses) or valid negatives (VN). For any set of R alternatives, then, there are three correct scores: AA, BB, and CC. However, there are also six possible error scores; A's called B, B's called A, A's called C, C's called A, B's called C, and C's called B. The error scores differ in the direction and magnitude of error. For instance, a C called A represents calling a non-relevant

response relevant while a B called A represents calling a moderately relevant response relevant, etc. The various types of correct responses and error responses are summarized below:

Positive Identifications (PI)

AA (relevant called relevant)

BB (moderately relevant called moderately relevant)

Valid Negatives (VN)

CC (not relevant called not relevant)

Total Correct Responses (CR)

PI + VN = CR (AA + BB + CC = CR)

Error Responses

AB (relevant called moderately relevant)

BA (moderately relevant called relevant)

CB (not relevant called moderately relevant)

BC (moderately relevant called not relevant)

AC (relevant called not relevant)

CA (not relevant called relevant)

Each type of correct response can be summed within as well as summed over problems. Both sums of scores by types within problems and over problems were used in the present study. However, unless otherwise specified; "PI score" means the AA + BB responses summed over problems, "VN score" means the CC responses summed over problems, and "CR score" means AA + BB + CC summed over problems. It might have been assumed

that the PI score and the VN score, representing correct responses, measure the same thing. However this was not assumed since it was not known whether ability to eliminate "poor" alternatives is the same as ability to identify "good" alternatives.

Only four of the seven problems used could be dealt with in the manner described above. Of the remaining problems one was responded to by rank ordering, and two were left as free written response problems. Problem descriptions follow:

Problem A. Problem A requires that the subject examine 10 examples in long division solved by a student in grade 5. On the basis of this examination the subject is asked to make tentative judgments concerning what actions might be taken to remedy the errors made by the student. There are 14 alternative actions to be judged. Each is judged to be either (A) a sufficient action (treated as totally relevant) or (B) a necessary, but not sufficient action (treated as moderately relevant) or (C) an unnecessary action (treated as not relevant). The a priori correctness of response for any given alternative is based on the pattern of error in the long division examples, which is systematic in one class of examples and random in the others. The pattern of error does not permit a judgment of sufficiency for any one of the alternatives, although a combination of four alternatives does lead to sufficiency, hence there are four correct B responses and 10 correct C responses and no correct A responses in this problem.

Problem B. Problem B requires the subject to judge the relevance

(A, B, or C) of an exercise from a third grade arithmetic text to 15 objectives of arithmetic instruction. The objective of the exercise as given by the authors of the text was used as the relevant alternative while the not relevant alternatives were selected on the basis of contradiction with the structure or content of the exercise or of having no referent in the exercise. The moderately relevant alternatives were determined on the basis of whether they contained references to at least one of the operations required of the student in the exercise. There are one A alternative, four B alternatives, and 10 C alternatives in Problem B.

Problem E. Problem E contains 12 arithmetic examples and six problems at fourth grade level. As in Problem A, errors occur both systematically and randomly in this exercise. The task for the teacher is to determine which alternatives have a bearing on errors that should be made the focus of an interview with the pupil making them, which alternatives have a bearing on errors that are peripheral but might be included and which alternatives are irrelevant to the interview. The systematic errors determine the one relevant alternative, the random errors determine the four "might be included" or moderately relevant alternatives, while the 10 not relevant alternatives have no bearing on the errors actually made.

Problem H. Problem H requires the subject to judge the relevance of an exercise used in the "middle" grades to 10 objectives of arithmetic instruction. Alternatives were judged relevant, moderately relevant or not relevant by the same method as used in Problem B. In Problem H there

are two relevant alternatives, two moderately relevant, and six not relevant.

Problem F. Problem F requires the subject to place seven long division examples in order according to difficulty for middle grade students. Difficulty in this set of problems was determined on an empirical basis following Brueckner and Grossnickle (2), (p. 284). This problem is an interesting one in that it contains "noisy" attributes which appear to determine difficulty but in fact do not. Correctly ranked examples were summed in the PI score on this problem while errors were arbitrarily placed in the BC error category from which they could be removed conveniently during detailed analysis.

Problems D and G. Problem D requires the subject to state the "meanings" of subtraction as principles to serve as guides to students. Problem G requires the subject to state the "meanings" of division. Scoring was based on the standard "meanings" of these processes. Response exemplars for use in scoring were obtained from the responses of teachers to earlier, but identical, forms of these two problems. A total score of four was assigned to each problem.

Determination of random response scores. The conversion of free response problems to problems which could be objectively scored created the possibility that subjects could obtain scores on the basis of random responses. This possibility in turn suggested that a statistical "control group" (representing the scores that would occur if no learning, i.e. selective

responses, took place) could be defined by the random response distribution. Such a distribution is a convenient reference point with which to assess differences between groups with varying exposures to opportunities to acquire instrumental responses and serves as well as a means of defining proficiency since the least proficient problem solvers would ipso facto respond at random while the most proficient would respond least randomly or most selectively.

The random response distribution for the PI scores of problems A, B, E, and H was calculated on the basis of the binomial theorem, yielding a mean score of 6.00 and a variance of 4.00 when N = 100. For Problem F the mean is 1.06 with a variance of .99 when N = 100. This value was obtained by drawing fifty sets of ranks from a table of random numbers, doubling the number of occurrences of one correct rank, two correct ranks, etc., then taking the mean and variance of the resulting distribution. The appropriate values in this distribution may be checked by the equation:

$$pm = \frac{e^{-1}}{m!}$$

where p. is the probability of a match, $e^{-1} = .3675$, and m is the number of matches (10, p. 67). In Problem F, m varies from 0 to 7. For the two free response tasks, the probability of a correct response by random responses was estimated to be close to zero. The random response mean of the PI score was thus determined to be 7.06 and the variance 4.99, when N = 100. The family of curves to which the random response distribution belongs was not determined, but was assumed to be approximately normal; utilization of this assumption yielded a score value of 11.16 for the 5%

point when N = 100. This is to say that one would expect a PI score of approximately 11 to arise 5% of the time by chance alone. This does not mean that a score of 11 is necessarily a chance score, but it does mean that as scores fall at or below 11 there is decreasing confidence in the score as one obtained by selective responding.

The value of the mean and variance of the random response distribution for the CR score was determined by the same procedures as above. These procedures yielded a mean of 19.06 with a variance of 12.99 and standard deviation of 3.60.

Sampling Procedures

In order to test the problems against the several validating criteria earlier set forth and to obtain data on reliability, samples were drawn from several pools.

Teacher sample. Teachers were drawn from two Indiana school systems. In the first, a small consolidated system with 45 teachers in grades 3-6, 41 teachers participated. In the second, a large consolidated system with 98 teachers, grades 3-6, 95 teachers participated. These systems were used primarily because they provided teachers who teach in rural and village schools, teachers who teach in small city or "town" (population about 6,000) schools, and teachers who teach in city (population about 38,000) schools, as well as providing teachers from systems of different size. This sample seemed not only to be a good cross-section of teachers which was necessary for validation purposes, but also permitted analysis for differences in

in teacher performance according to location.

Non-teacher sample. The non-teachers of arithmetic, grades 3-6, were drawn from four pools: (i) nurses who had returned to Indiana University to complete the B.S. or advanced degrees; (ii) secondary teachers, businessmen, nurses, and others who were enrolled in an elementary statistics class at an Indiana University extension center; (iii) members of the Tri Kappa Sorority in an Indiana city; and (iv) persons in the age group 23-60 who had returned to Indiana University to work toward teaching certificates, but who had no teaching experience and no formal course work in arithmetic methods. The sample from these four pools totaled 41 persons.

Preparatory teacher sample. Preparatory teachers were used to estimate the stability (between occasions reliability) of the problems and as one estimate of the within-occasions reliability. In addition they were used: (i) to determine the relevance of arithmetic methods, as a treatment, to performance; (ii) as a control group from which to determine the effects of teaching experience; and (iii) as a relevant population for testing the hypotheses bearing on intelligence, personal values and attitudes toward children in relation to problem solving performance. The advantage of using preparatory teachers to test the latter hypotheses lies in being able to maintain control over teaching experience.

The sample of preparatory teachers was composed of three pools. The first two pools consisted of students enrolled in an arithmetic methods course and who subsequently were assigned either to grades 3-6 for student

teaching, or to kindergarten, grade 1 or grade 2 (K-2). The first pool was assessed before the arithmetic methods course formally began and approximately 7 months before student teaching was begun. For convenience the first group is designated bams 3-6 (before arithmetic methods, subsequently assigned to 3-6), and the second group bams K-2.

The second pool was assessed after arithmetic methods but before student teaching. The members of this pool are designated aams 3-6 (after arithmetic methods, subsequently assigned to 3-6) and aams K-2.

The third pool consisted of members of an arithmetic methods course who subsequently were assigned to student teaching in either grades 3-6 or K-2 but who were assessed after both methods and student teaching, and those persons who had completed arithmetic methods at extension centers, who had completed student teaching, and who had returned to the main Indiana University campus for a workshop just prior to graduation, at which time they were assessed. This pool was divided according to whether student teaching was done in 3-6 or K-2. These groups are designated ast(1) 3-6 and ast(1) K-2.

Both the aams 3-6 and the aams K-2 group were assessed again after they returned from student teaching, an interval of about 4 months. For the second assessment the designation is ast(2) 3-6 and ast(2) K-2, with the (2) designating the number of times the group was assessed.

In order to obtain an estimate of the relevance of arithmetic methods as a treatment the differences between matched pairs from pool 1 and pool 2 were examined. In order to obtain stability coefficients and make allowance

for possibly relevant intervening treatments the performances of aams 3-6 with ast(2) 3-6 were correlated. To control for practice effects the differences between ast(1) 3-6 and ast(2) 3-6 were observed.

The number of subjects in each group may be observed in Table 1.

TABLE 1. NUMBER OF SUBJECTS IN EACH GROUP.

Teachers	Non-Teachers	Preparatory Teachers		
		Pool 1	Pool 2	Pool 3
136	41	80*	63**	52

bams		aams		ast(1)	
3-6	K-2	3-6	K-2	3-6	K-2
31	45	28	25	25	27

ast(2)	
3-6	K-2
28	25

*Four subjects were lost from this pool before student teaching assignments were made and were discarded in sub-group comparisons.

**Ten subjects were lost from this pool after the first assessment and were discarded in sub-group comparisons.

Method of obtaining data on independent variables. For preparatory teachers, The American Council on Education Psychological Examination (ACE) (13) was used as the measure of intelligence, and the Cooperative English Test, Test C_2, Reading Comprehension (Coop C_2) (9) as the measure of reading comprehension. These scores were obtained from the Indiana University

Test Bureau. Scores were not available for all subjects. The Minnesota Teacher Attitude Inventory (MTAI) (8) and the Allport-Vernon-Lindsey Study of Values (1) were administered to groups ast(2) 3-6, ast(2) K-2, ast(1) 3-6 and ast(1) K-2 two days before the end of their final semester. Scores on the latter instruments were not available for those who departed from the campus early or who were ill at the end of the semester.

Data bearing on the various characteristics of teachers and non-teachers was obtained by questionnaire. The reliability of the questionnaire was not computed, but the data obtained by this method coincided with information obtained from independent sources.

RESULTS

The Differentiation of Teachers from Non-Teachers

Sample characteristics. The possible significance of differences between teachers and non-teachers rests partly on the comparability of the samples in respects other than professional training and teaching experience. Comparisons of the samples with respect to years of higher education, number of courses in mathematics, age and sex are shown in Tables 2, 3, 4, and 5 respectively.

TABLE 2. NUMBER OF TEACHERS AND NON-TEACHERS BY YEARS OF HIGHER EDUCATION

Group		Number of Years of Higher Education					
		fewer than 4	4	more than 4 less than 5	5	more than 5	N
Teachers	n	20	70	16	26	4	136
	%	14.7	51.4	11.8	19.1	3.0	100
Non-Teachers	n	9	16	5	4	7	41
	%	22.0	39.0	12.2	9.8	17.0	100

TABLE 3. NUMBER OF TEACHERS AND NON-TEACHERS BY NUMBER OF COURSES IN MATHEMATICS.

Group		Number of Courses in Mathematics				
		1-2	3-4	5 or more	Unknown	N
Teachers	n	44	62	28	2	136
	%	32.4	45.6	20.6	1.4	100
Non-teachers	n	12	12	16	1	41
	%	29.3	29.3	39.0	2.4	100

TABLE 4. NUMBER OF TEACHERS AND NON-TEACHERS IN SIX AGE CATEGORIES.

Group		Age Group						
		22 and below	23-29	30-39	40-49	50-59	60 +	N
Teachers	n	9	38	16	35	35	3	136
	%	6.6	27.9	11.8	25.7	25.7	2.3	100
Non-teachers	n	0	18	16	6	1	0	41
	%	0	44.0	39.0	14.6	2.4	0	100

TABLE 5. NUMBER OF MEN AND WOMEN IN THE TEACHER AND NON-TEACHER SAMPLES.

Group		Men	Women	N
Teachers	n	29	107	136
	%	21.3	78.7	100
Non-teachers	n	18	23	41
	%	44.0	56.0	100

The sample of non-teachers is slightly more educated (non-teacher $\overline{X}$ = 4.27, teacher $\overline{X}$ = 4.15), somewhat younger, has had somewhat more mathematics and contains proportionally more men than the teacher sample. In view of subsequent results, whether any of these differences are of any practical significance is doubtful.

Validity of criterion scores. Three scores, PI, VN, and CR, were initially tested to determine which yielded the greatest differences between teachers and non-teachers. The results for PI, VN, and CR may be observed in Tables 6, 7, and 8 respectively.

TABLE 6. MEANS AND VARIANCES OF PI SCORE AND t VALUE: TEACHER VS. NON-TEACHERS.

Group	N	s^2	F	$\overline{X}$	t	p
Teachers	136	18.13		14.69		
			1.56 (ns)		5.912	.001
Non-teachers	41	11.60		10.88		

TABLE 7. MEANS AND VARIANCES OF VN SCORE AND t VALUE TEACHERS VS. NON-TEACHERS.

Group	N	s^2	F	$\overline{X}$	t	p
Teachers	136	39.98		18.49		
			1.07 (ns)		.890	ns
Non-teachers	41	42.83		17.46		

TABLE 8. MEANS AND VARIANCES OF CR SCORE AND t VALUES: TEACHERS VS. NON-TEACHERS.

Group	N	s^2	F	$\bar{X}$	t	p
Teachers	136	65.35	1.16 (ns)	33.18	3.013	.01
Non-teachers	41	75.85		28.34		

Since PI + VN = CR, the significance of the CR score stems principally from its PI component. The PI score may thus be viewed as the continuous score in which greatest confidence, with respect to validity, may be placed. However, during scoring and analysis it was noted that the PI score is subject to two weaknesses. First, it cannot show how consistently a subject performs, i.e. whether he does reasonably well on all problems and gets a high score by this means or whether he does well on two or three problems but poorly on the others and gets a reasonably high score by this means. Second, the PI score can be biased by response preferences, i.e. continuously choosing a particular response, such as "B".

To take consistency of performance and response preferences into account, an additional criterion, the "consistency criterion", was developed. In developing the consistency criterion, Problem G, to which few persons correctly responded, was deleted. Subjects were categorized according to performance on the remaining six problems. Those who scored at least one point above the most probable chance score on both the PI and VN score on four of the six problems and who scored on the most probable chance level

on only the PI or only the VN score, on any remaining two problems, were placed in a "high consistency" group. Those who scored at the most probable chance level or below on both the PI and the VN score on at least four of the six problems were placed in a "low consistency" group. The remaining subjects were placed in a middle group. This method of grouping not only arrayed performance according to consistency, but also eliminated from the high group any subject who showed a preference over several problems for a particular response, e.g. B. A consistent preference for a B response, for instance, while inflating the PI score, strictly entails a VN score below the most probable chance score on each problem on which a "B" response preference is shown.

The validity of the consistency criterion with respect to the differentiation of teachers from non-teachers may be observed in Table 9.

TABLE 9. TEACHERS VS. NON-TEACHERS ACCORING TO CONSISTENCY OF PERFORMANCE.

	Consistency Groups			
Group	High	Middle	Low	N
Teacher	32	68	36	136
Non-teacher	2	20	19	41

$X^2 = 9.73$, df = 2, $p < .01$

On the basis of the PI score and the consistency criterion, the inference seems warranted that teachers, as a group, not only make significantly more positive identifications or goal achieving responses, but also yield a higher

proportion of persons whose performance is consistently "good" over several problems and a smaller proportion of persons whose performance is consistently "poor" over several problems than is the case with non-teachers.

Sensitivity of the Problems to Differences in Teaching Experience

Teaching experience and age are correlated. It is desirable, therefore, to hold experience constant and examine for the independent effect of age in making assertions about teaching experience. This is done at a later point in this section. In the teacher sample drawn, changes in the size of the institutions (as indexed by the size of the student body) where teachers were prepared and/or graduated also accompany increases in teaching experience. This effect may be observed in Table 10. In order to extract the effects of teaching experience, it is therefore also desirable to hold size of graduating institution constant.

Tables 11 and 12, based on analysis of the PI score, show the F and t values respectively for teachers with 0, 1-3, 4-10, and 11-25 years of teaching experience who hold degrees from institutions larger than 5000 students. The teachers with 0 experience are the ast(1) 3-6 group which was measured only one time and which had been prepared to teach grades 3-6. All teachers who prepared at large institutions but had not taken at least Bachelor degrees were excluded. The latter policy reduced the number of teachers in the 26 - 42 group to three, and they were discarded as an insufficient sample. The mean PI score for the three persons discarded was 14.66

TABLE 10. PROPORTION OF TEACHERS AT FOUR LEVELS OF EXPERIENCE COMING FROM SMALL, MEDIUM, AND LARGE PREPARATORY INSTITUTIONS.

Experience Level*	Institution Size (number of students) 0 - 999		1000 - 4999		5000 & above		
	N	%	N	%	N	%	N
1- 3 years	2	5.26	16	42.11	20	52.63	38
4-10 years	10	28.57	9	25.72	16	45.71	35
11-25 years	15	38.46	9	23.08	15	38.46	39
26-42 years	7	30.43	9	39.14	7	30.43	23
Totals	34		43		58		135

* not available for one teacher

TABLE 11. ANALYSIS OF VARIANCE FOR PI SCORES OF TEACHERS WITH 0, 1-3, 4-10, AND 11-25 YEARS EXPERIENCE.

Source of Variation	df	ss	Mean Square
Groups	3	338	112.67
Within	70	1181	16.87
Totals	73	1519	
Groups ÷ Within = 6.68,		p .02	
F for extreme variances = 2.49,		df = 18 and 14,	p .05 .10

TABLE 12. MEANS AND VARIANCES OF PI SCORES, AND t VALUES FOR TEACHERS WITH 0, 1-3, 4-10, AND 11-25 YEARS EXPERIENCE.

Group	n	s^2	$\bar{X}$	t	p
Student Teachers ast(1) 3-6	35	13.58	11.56		
				3.399	.01
Teachers 1-3 yrs. exp.	20	27.36	16.25		
Teachers 4-10 yrs. exp.	15	11.00	16.53		
				1.760	n.s.
Teachers 11-25 yrs. exp.	14	13.92	14.22		

On the basis of the data in Table 12 it seems probable that the asymptote of the performance curve is reached during the very early years of experience. However, exactly where the asymptote lies and what individual differences exist in rate of attainment of peak performance cannot be adequately determined from cross-sectional data. In addition to suggesting that the asymptotic level of performance is reached early, the data in Table 12 also suggest that teaching experience is a variable relevant to performance. However, an examination of the effects of teaching experience in relation to graduates of small colleges, as presented in Table 13, casts some doubt on whether teaching experience alone will contribute to increases in performance. In Table 13, teachers with 1-3 and 26-42 years of experience were excluded because of insufficient numbers. All teachers held degrees from small institutions.

TABLE 13. MEANS AND VARIANCES OF PI SCORE FOR TEACHERS WITH 0 YEARS EXPERIENCE FROM LARGE PREPARATORY INSTITUTIONS, AND TEACHERS WITH 4-10 AND 11-25 YEARS EXPERIENCE FROM SMALL INSTITUTIONS.

Group	n	College Size	s^2	$\bar{X}$
Student Teachers ast(1) 3-6	25	above 5000	13.58	11.56
Teachers 4-10 yrs. exp.	7	below 1000	17.00	11.42
Teachers 11-25 yrs. exp.	10	below 1000	16.67	12.50

On the basis of the data in Tables 12 and 13 it seems probable that variables associated with the size of the institution from which the teacher graduated are closely related to the acquisition of arithmetic teaching skills during experience. What these variables are and how they bear on the acquisition of arithmetic teaching skills was not investigated in the present study.

It is possible that what has been attributed to teaching experience thus far might in fact be attributable to age since age and experience are correlated. To check this possibility, teachers who were graduated from institutions larger than 999 students, who were between 23 and 29 years of age, and who had had 1-7 years experience were compared to teachers who were graduated from institutions larger than 999 students, who were between 30 and 49 years of age, and who had had 1-7 years experience. Six of the teachers in the latter group were 30-39 years old, and had a mean score of

15.33. Six were 40-49 years old and had a mean score of 14.50. These two groups were pooled in the analysis shown in Table 14. From the data in Table 14, it seems very probable that up to age 50 at least, age alone affects performance comparatively little.

TABLE 14. MEANS AND VARIANCES OF PI SCORE, AND t VALUE FOR TEACHERS OF 1-7 YEARS EXPERIENCE, AGE 23-29, AND 30-49.

Group	n	s^2	F	$\bar{X}$	t	p
Age 23-29	43	19.64	1.02 (ns)	15.86	.616	< .50
Age 30-49	12	20.09		14.92		

In addition to using the PI score criterion, the main results for teaching experience were also checked on the consistency criterion for teachers with 0 and 1-3 years experience with college size held above 5000. The results may be observed in Table 15.

TABLE 15. TEACHERS WITH 0 YEARS EXPERIENCE VS. TEACHERS WITH 1-3 YEARS EXPERIENCE ON CONSISTENCY OF PERFORMANCE.

Experience Group	Consistency Groups: High	Middle	Low	N
Student Teacher ast(1) 3-6	4	12	9	25
Teachers 1-3 yrs. exp.	10	8	2	20

$X^2 = 8.81$, df = 2, $p < .02$

From these results and those in Table 12, it is apparent that neophyte teachers not only obtain significantly higher PI scores, but also solve the teaching problems with greater consistency than do preparatory teachers at the end of training.

Relationships Between Problem Solving Performance and Selected Independent Variables

Three dependent variable criteria were used in identifying relationships between independent variables and problem performance; the continuous PI score, the consistency criterion, and a third criterion based on 3 PI cutting scores. The cutting scores were determined by reference to the teacher, non-teacher, and random responses distributions. Persons who fall below score point 11 on the PI distribution approach the random response level and clearly show few indications of skill in solving problems of the type used. Score point 15 on the PI distribution divides the teacher sample approximately in half while score point 18 on the PI distribution cuts the lower three quarters of the teachers. The percentages of teachers and non-teachers cut at each of these scores may be observed in Table 16.

TABLE 16. PERCENTAGE OF TEACHERS AND NON-TEACHERS AT OR ABOVE EACH OF THREE PI SCORES.

Group	PI Scores 11	15	18
Teachers	75% above	51.47% at or above	25.00% at or above
Non-teachers	46.43% above	9.76% at or above	2.44% at or above

Intelligence. The relationship between intelligence and problem solving performance was investigated using the continuous PI score criterion with two groups, bams K-2 + bams 3-6 and aams 3-6 + aams K-2 and the consistency criterion for ast(1) 3-6 + ast(2) 3-6 + ast(1) K-2 + ast(2) K-2. Scores were not available for all subjects in each group.

The Pearson's r between ACE raw score and the PI score for bams K-2 + bams 3-6 is .29 ($n = 57$, $p < .05$) and for aams K-2 + aams 3-6 is .50 ($n = 46$, $p < .01$). Results of the comparison of ACE raw scores of subjects with high consistency in performance versus subjects with low consistency in performance by the t test may be observed in Table 17.

TABLE 17. MEANS AND VARIANCES OF ACE RAW SCORES, AND t VALUE, HIGH CONSISTENCY VS. LOW CONSISTENCY STUDENT TEACHERS.

Consistency Group	n	s^2	F	$\bar{X}$	t	p
High	12	553.33	1.39 (ns)	116.00	1.993	>.05<.10
Low	13	397.00		98.67		

While it seems likely that the differences in the means of the high and low consistency groups in ACE score is a true difference, full confidence on this matter awaits a replication of the differences in a subsequent study.

It should be noted that the ACE scores used to obtain the above results were taken when the students were admitted to college and that the r's given indicate the ability of the ACE to predict PI scores obtained

approximately three years later.

MTAI and Study of Values. The relationships between MTAI scores, Study of Values scores, and the PI score were computed independently for the ast(1) K-2 + ast(1) 3-6 group and the ast(2) K-2 + ast(2) 3-6 group, thus yielding a cross-validation of each r. The r's were computed using only those subjects in each group for whom both MTAI and Study of Values scores were available. The results may be observed in Table 18.

TABLE 18. CORRELATIONS BETWEEN THE MTAI, THE STUDY OF VALUES, AND THE PI SCORE FOR TWO GROUPS OF STUDENT TEACHERS.

Scale	Group	
	ast (1) (40 df)	ast (2) (38 df)
MTAI	.42**	.33*
Theoretical	.16	.23
Economic	-.07	-.27
Aesthetic	.14	.32*
Social	-.22	.41**
Political	.06	.26
Religious	.32	.21

*Significant at p = .05
**Significant at p = .01

MTAI and Study of Values scores were also compared for high and low groups on the consistency criterion. To obtain a reasonable number of subjects

in the high and the low groups, all student teachers ast(1) + ast(2) were pooled. MTAI scores but not Study of Values scores were available for two subjects in the high group and two subjects in the low group, thus changing the N between the segment of Table 19 showing MTAI scores and those segments showing the Study of Values scores.

From the data in Tables 18 and 19 it is apparent that responses to the MTAI are consistently related to problem solving performance among student teachers. Whether the same is true for the teacher population awaits further study. The failure of the various scales on the Study of Values to hold consistent relationships to performance in the present samples suggests that there is at best only a somewhat tenuous relationship between problem solving performance in arithmetic and the measured values of student teachers. It is of course possible that the values of experienced teachers are consistently related to problem solving performance and this possibility should be explored before the Study of Values is dismissed as irrelevant.

Teaching location. Of the 136 subjects in the teacher sample, 25 had originally been employed by county superintendents to teach in county schools. The county schools in which these 25 teachers taught were later brought into consolidated metropolitan districts. Of the remaining 111 teachers, 74 were employed by a city superintendent to teach in metropolitan schools, while 37 were employed by a "town" superintendent to teach in the small metropolitan district centering in that particular town. The mean PI scores of these three groups of teachers were examined for differences. The results may be observed in Table 20.

TABLE 19. MEANS AND VARIANCES AND t VALUES OF MTAI AND STUDY OF VALUES SCORES FOR STUDENT TEACHERS OF HIGH AND LOW CONSISTENCY.

Group	n	s^2	MTAI F	$\bar{X}$	t	p
Low	20	45.93		59.61		
			1.19 (ns)		8.147	<.001
High	18	54.84		40.80		
			Theoretical Scale			
Low	18	35.12		41.22		
			1.68 (ns)		2.751	<.02
High	16	59.00		34.69		
			Economic Scale			
Low	18	48.18		38.50		
			1.11 (ns)		.350	ns
High	16	43.53		37.69		
			Aesthetic Scale			
Low	18	53.41		39.72		
			1.44 (ns)		1.757	ns
High	16	37.07		43.81		
			Social Scale			
Low	18	35.76		38.83		
			1.53 (ns)		.089	ns
High	16	54.80		38.63		
			Political Scale			
Low	18	34.12		38.28		
			1.38 (ns)		1.127	ns
High	16	27.73		36.19		
			Religious Scale			
Low	18	77.76		42.83		
			2.61 (ns)		2.306	<.05
High	16	29.60		48.56		

TABLE 20. ANALYSIS OF VARIANCE OF PI SCORES FOR COUNTY, TOWN, AND CITY TEACHERS.

Source of Variation	df	ss	Mean Square
Groups	2	31.00	15.50
Within	133	2416.00	18.17
Totals	135	2447.00	
F for extreme variances = 1.16,	df = 73 and 36,	p .10	

Since the PI score criterion yielded no differences between groups, the method of grouping teachers was changed for the test of location of teaching on the consistency criterion. Town and county teachers in the small metropolitan district were pooled and city and county teachers in the large metropolitan district were pooled, and the number of teachers falling in the high consistency category and low consistency category from these pools compared. The results may be observed in Table 21.

TABLE 21. CONSISTENCY OF PERFORMANCE OF TEACHERS FROM A SMALL CONSOLIDATED SYSTEM VS. TEACHERS FROM A LARGE CONSOLIDATED SYSTEM.

System Size	Consistency Group High	Low	N
Small	9	14	23
Large	23	22	45
x^2 = .91,	df = 1,	p .30	

Size of teacher preparatory institution. Preparatory institutions were classified

into three groups (0-999, 1000-4999, 5000 and over) according to the number of students in the student body as reported in the 1957-58 Education Directory (16). First, the number of teachers in each category of institution size falling at each of four PI score intervals was observed, and second the number in the same categories of institution size falling in the high and low groups on the consistency criterion was observed. The results may be examined in Tables 22 and 23 respectively.

TABLE 22. NUMBER OF TEACHERS PREPARED AT INSTITUTIONS OF SIZE 0-999, 1000-4999, AND 5000 AND ABOVE AT EACH OF FOUR PI SCORE INTERVALS.

Institution Size	PI Score Interval				
	11 & below	12 - 14	15 - 17	18 & above	N
0-999	16	7	4	7	34
1000-4999	9	13	10	11	43
5000 & above	9	12	22	16	59
$x^2 = 16.26$,	df = 6,	p < .02			

TABLE 23. NUMBER OF TEACHERS PREPARED AT INSTITUTIONS OF SIZE 0-999, 1000-4999, AND 5000 AND ABOVE IN THE HIGH AND LOW CONSISTENCY GROUPS.

Institution Size	Consistency Group		
	High	Low	N
0-999	5	13	18
1000-4999	7	13	20
5000 & above	20	10	30
$x^2 = 8.57$,	df = 2,	p < .02	

On the basis of Tables 22 and 23 it appears that the size of the institution at whicha teacheris preparedisa significant variable. However, the size of the institution at which a teacher prepares does not seem intrinsically related to performance, rather it would appear to function as an index of other variables which presumably are intrinsically related. What these variables are remains to be investigated in subsequent studies.

Reading comprehension. The relationship between reading comprehension as measured by the Coop C_2 and problem solving performance was investigated using the PI score criterion with two groups, bams K-2 + bams 3-6 and aams 3-6 + aams K-2, and the consistency criterion for ast(1) + ast(2). Scores were not available for all subjects.

The r between Coop C_2 raw scores and the PI score for bams K-2 + bams 306 is .20 ($n = 57$, $p < .05$) and for aams 3-6 + aams K-2, .45 ($n = 46$, p .01). Results of the comparisons of subjects with high consistency versus subjects with low consistency by the t test may be observed in Table 24.

TABLE 24. MEANS AND VARIANCES OF COOP C_2 RAW SCORES, AND t VALUE HIGH CONSISTENCY VS. LOW CONSISTENCY STUDENT TEACHERS.

Consistency Group	n	s^2	F	$\bar{X}$	t	p
High	13	62.50	1.86 (ns)	59.80	2.068	>.05<.10
Low	12	33.55		54.08		

On the basis of the overall results stemming from the use of the Coop C_2 there appears to be little doubt that reading comprehension plays some role in problem performance. However, the Coop C_2 and the ACE correlate closely in the undergraduate samples (.74 to .84) so that reading comprehension and intelligence are intrinsically confounded.

Grade level. The relationship of the grade level at which a teacher teaches to problem performance was greatly complicated by the fact that most of the teachers in the sample either were teaching or had taught more than one grade level. To take into account the range of grade levels in which teaching experience had occurred, three groupings of teachers were made. Group 1 included teachers who were teaching grade 3 and/or grade 4 and who previously had taught only these grades either separately or in combination. Group 2 includes teachers who were teaching grade 5 and/or grade 6 and who previously had taught these grades either separately or in combination. Group 3 includes teachers who were teaching one grade or combination of two grades, (3 - 4 or 5 - 6), but who had before taught all grades 3-6. These classifications exclude teachers who had taught only at the extremes, i.e. 3 and 6. There were 12 such teachers. The number of teachers in each of the first three groups at four PI score levels were then compared. The results may be observed in Table 25.

TABLE 25. NUMBER OF TEACHERS IN FOUR PI SCORE INTERVALS WHEN CLASSIFIED BY GRADE LEVELS TAUGHT.

Grade Levels	PI Score Interval				N
	11 & below	12 - 14	15 - 17	18 & above	
3 - 4	15	7	11	12	45
5 - 6	8	9	9	8	34
3 - 6	9	15	11	10	45
x^2 = 3.80,		6 df,	p .70		

It is apparent from the results in Table 24 that grade level, as classified, makes no difference in performance. An additional check on grade level as a variable was done as part of a two-way analysis of variance and may be observed on page 43.

Arithmetic methods as a treatment. To assess arithmetic methods as a treatment of possible relevance to problem solving performance 30 students from the bams K-2 + bams 3-6 groups were matched for sex and ACE score with 30 students from the aams K-2 + 3-6 groups. All students were between 20 and 22 years old and the size of preparatory institution was held constant. The PI score was used as the criterion score. Results may be observed in Table 26.

TABLE 26. MEANS AND VARIANCES OF PI SCORE FOR MATCHED PREPARATORY TEACHERS BEFORE AND AFTER ARITHMETIC METHODS.

Group	n	s^2	$\bar{X}$
Before Methods	30	11.03	11.93
After Methods	30	17.70	11.78

While the amount of control in matching independent groups is much less than when each subject is matched with himself, it is clear that problem solving performance as measured is not sensitive to whatever changes result from exposure to an arithmetic methods course.

Recency of exposure to arithmetic methods. For the initial analysis two groups were used: teachers who had taken an arithmetic methods course within the past nine years and teachers who had taken methods 10 or more years ago. The frequency of teachers in each of these groups in four PI score intervals may be observed in Table 27. Data were not available for 12 teachers.

TABLE 27. NUMBER OF TEACHERS WITH ARITHMETIC METHODS WITHIN THE PAST NINE YEARS VS. THOSE WITH METHODS 10 OR MORE YEARS AGO IN FOUR PI SCORE INTERVALS.

Years Since Methods	PI Score Interval				
	11 & below	12 - 14	15 - 17	18 & above	N
0-9	19	14	21	15	69
10 or more	13	11	14	17	55

$x^2 = 1.50$, df = 3, $p > .50$

Because the effects of arithmetic methods may be contingent not only on when they were taken, but also where they were taken, the interaction between recency of methods and size of institution at which they were taken was examined. To avoid great disproportionality between cells, teachers were re-grouped so that teachers having a methods course 10 years ago were removed from the 10 or more group and placed with the 0-9 group. Groupings with respect to institution size were also changed, only the institution at which the person last took his methods course was used, rather than size of graduating institution. These institutions were placed in only two classes, 0-2999 and 3000 and above. The results of the analysis using the foregoing classifications may be observed in Table 28.

TABLE 28. ANALYSIS OF VARIANCE OF PI SCORES: SIZE OF INSTITUTION AT WHICH METHODS WERE LAST COMPLETED BY YEARS SINCE COMPLETION.

Source of Variation	df	Sums of Squares		Mean Square
		Unadjusted ss*	Adjusted ss	
Years since methods	1	3.99	7.32	7.32
Institution size	1	23.46	26.79	26.79
Interaction	1	49.20	45.87	45.87
Within	120	1985.12		16.64
Totals	123	2061.77		

*The adjustment term for disproportionality is -3.33; it is subtracted from the ss for main effects and added to the ss for interaction.

On the basis of the data in Tables 27 and 28 it is apparent that neither the recency with which methods were taken nor the size of institution at which they were taken nor the interaction between is significant. Whether the same would be true in larger samples where more sensitive classifications can be used must remain for the time being a matter of speculation.

Number of courses in mathematics. A "mathematics" course means in the present study: plane geometry, first year algebra, second year high school algebra, solid geometry, trigonometry, college algebra, analytic geometry, calculus, and courses beyond calculus. General high school mathematics and special mathematics courses for elementary teachers were excluded. Defined in this way, the number of courses in mathematics among teachers were categorized as follows: (i) one or two math courses, usually meaning first year algebra and plane geometry; (ii) three or four math courses usually meaning in addition to first year algebra and plane geometry, second year algebra or college algebra, trigonometry or solid geometry; (iii) five or more courses, which predominantly included teachers with one or two mathematics courses in college together with considerable high school mathematics. Categorized in this way, courses in mathematics were run against four PI cutting score intervals. Information was not available for two teachers. Results may be observed in Table 29.

To make a more sensitive test of the possible significance of the grade level at which the teacher was teaching, number of mathematics courses he had taken, and to investigate the interaction between, a 2 x 2 analysis of variance

was done. For this analysis, all teachers from institutions smaller than 2000 were excluded. Of the remaining group only teachers who had had experience in either grade 3 and/or grade 4, or grade 5 and/or grade 6 were included. These groups were classified according to whether they had had a mimimum of mathematics (0-2 courses) or more than a minimum of such courses (3 or more courses). The results may be observed in Table 30.

TABLE 29. NUMBER OF TEACHERS WITH 1-2, 3-4, AND 5 OR MORE MATHEMATICS COURSES IN FOUR PI SCORE INTERVALS.

Number of Math Courses	PI Score Interval			
	11 & below	12 - 14	15 - 17	18 & above	N
1 - 2	11	11	14	8	44
3 - 4	17	15	15	15	62
5 or more	5	5	7	11	38
$x^2 = 4.853$,	df = 6,	p>.50			

TABLE 30. ANALYSIS OF VARIANCE OF PI SCORES: NUMBER OF COURSES IN MATHEMATICS BY GRADE LEVEL TAUGHT.

Source of Variation	df	Sums of Squares		Mean Square
		Unadjusted ss**	Adjusted ss	
Number of math courses	1	5.46	8.03	8.03
Grade level taught	1	0.91	3.48	3.48
Interaction	1	29.46	26.89	26.89
Within	47	770.00		16.38
Totals	50	805.83		

**The adjustment term for disproportionality is 2.57; it is added to the ss for main effects and subtracted from the ss for interaction.

None of the effects shown in Table 30 are significant, indicating once more that neither the number of math courses nor the grade level at which the teacher teaches, nor the interaction between are significant sources of variation in problem solving performance.

Sex. Differences in performance by sex were examined by the number of each sex falling at four PI score intervals, and by the number of each sex in the high consistency and low consistency groups. Results may be examined in Tables 31 and 32; none are statistically significant.

TABLE 31. NUMBER OF EACH SEX FALLING AT FOUR PI SCORE INTERVALS.

	PI Score Intervals				
Sex	11 & below	12 - 14	15 - 17	18 & above	N
Women	26	24	31	26	107
Men	8	8	5	8	29
$x^2 = 1.66$,	df = 3,	p>.50			

TABLE 32. NUMBER OF EACH SEX IN HIGH AND LOW CONSISTENCY GROUPS.

	Consistency Groups		
Sex	High	Low	N
Women	26	25	51
Men	6	11	17
$x^2 = .620$,	df = 1,	p>.30	

Error Scores

The error scores were examined primarily to note differences in response patterns on problems A, B, E, and H between teachers showing high consistency in performance, teachers showing low consistency, and non-teachers showing low consistency. In order to interpret the error scores it is convenient first to recognize the number of incorrect responses which would be made by the mean person responding randomly, and second the actual means attained by the three groups mentioned above. These data may be observed in Table 33.

TABLE 33. PATTERNS OF ERROR RESPONSES AMONG HIGH CONSISTENCY AND LOW CONSISTENCY TEACHERS, AND LOW CONSISTENCY NON-TEACHERS.

Type of Error Score	Errors by random response $\bar{X}$	Errors, high consistency teachers $\bar{X}$	Errors, low consistency teachers $\bar{X}$	Errors low consistency non-teachers $\bar{X}$
AC	1.33	.13	.22	.53
AB	1.33	.97	1.64	1.26
BA	4.66	3.03	5.16	5.47
BC	4.66	2.28	2.19	2.63
CA	12.00	2.84	8.72	8.63
CB	12.00	9.53	12.89	12.84

In examining Table 33, it may first be noted that no group seems to respond strictly at random, although both low consistency teachers and low consistency non-teachers approach the mean random response level more

nearly than high consistency teachers. A second aspect of Table 33 that is of interest is the tendency of both low consistency teachers and low consistency non-teachers to respond "upward." "Upward" responding means to attribute more relevance to a particular alternative than it actually bears. For instance, calling a C alternative either B or A, or a B alternative A is "upward" responding. If the means of the CB + CA + BA responses of the three groups are compared, substantial differences may be noted, being 15.40 for high consistency teachers, 26.67 for low consistency teachers and 26.94 for low consistency non-teachers.

The "downward" responses of the three groups, however, show greater homogeneity. The mean of AC + BC + AB for the high consistency teachers is 3.38, for the low consistency teachers 4.05, and for low consistency non-teachers 4.42.

There are 50 alternatives on which "upward" responses can be made, and 18 alternatives on which "downward" responses can be made. Accordingly, the mean high consistency teacher responds upward 30.80% of the time and downward 18.70% of the time, the mean low consistency teacher responds upward 53.34% of the time and downward 22.50% of the time, and the mean low consistency non-teacher responds upward 53.84% of the time and downward 24.56% of the time. These data indicate that persons who show low consistency in performance tend to overestimate the relevance to problem solution of many of the alternatives presented. Why they overestimate the relevance of alternatives was not explored in the present study, but might be worthy of examination in subsequent studies.

Inter-problem, Inter-score Relationships

The relationships of the PI score for each problem to the PI score of each of the remaining six problems and to each of the principal scores was computed for the consolidated town-county teachers, and the consolidated city-county teachers providing two independent estimates of the inter-problem, inter-score correlations. The matrices may be observed in Tables 34 and 35 respectively.

The data in Tables 34 and 35 suggest that the problems are reasonably independent of each other, with the exception of Problems D and G. However, the PI score for each problem has a restricted range resulting in small variances and consequently a distinct possibility of smaller r's than would be the case with larger variances.

Reliabilities

The reliabilities computed for the problems are of two types: those that give the reliability of the tasks within a single occasion, and those that indicate their reliability between occasions. The within-occasions reliabilities were run for several groups, the between occasions for only one group.

The within-occasions reliabilities were computed by the split-half method corrected for test length. Each problem was divided on the basis of an item analysis, so that equally difficult irrelevant alternatives, equally difficult moderately relevant and equally difficult relevant responses were paired in that problem. In the instances where there was only one relevant alternative in a problem, it was paired with itself. This procedure produced two parallel

TABLE 34. INTER-PROBLEM, INTER-SCORE CORRELATIONS FOR 41 TEACHERS IN THE CONSOLIDATED TOWN-COUNTY SYSTEM.

	Problems							Scores		
	A	B	E	F	H	D	G	PI	VN	CR
Problems										
A		-.24	-.15	-.20	.28	.30*	.15	.32*	.09	.23
B			.10	.07	-.07	.06	.10	.30*	.21	.29
E				.31*	.29	.13	-.11	.52**	.07	.31*
F					-.07	-.08	-.03	.51**	.17	.37*
H						.24	.07	.44**	-.08	.17
D							.43**	.60**	.44**	.61**
G								.41**	.34*	.44**
Scores										
PI									.41**	.78**
VN										.89**

*Significant at p = .05
**Significant at p = .01

TABLE 35. INTER-PROBLEM, INTER-SCORE CORRELATIONS FOR 95 TEACHERS IN THE CONSOLIDATED CITY-COUNTY SYSTEM.

	Problems						Scores		
	A	B	E	F	H	D	G	PI	VN	CR
Problems										
A		.16	.25	-.01	.22*	.09	.17	.51**	.10	.35**
B			.00	.08	.00	.11	.12	.44**	.25*	.39**
E				.14	.23*	.08	.02	.49**	.17	.40**
F					.00	.12	.02	.59**	.23*	.47**
H						-.03	.04	.32**	.14	.26*
D							.31**	.47**	.18	.36**
G								.44**	.02	.25*
Scores										
PI									.33**	.76**
VN										.85**

*Significant at p = .05
**Significant at p = .01

halves for the responses to each problem. Summing halves over problems produced two parallel forms for the CR score. When the paired irrelevant alternatives were dropped out, two parallel forms containing only alternatives counted in the PI score remained. Reliabilities were computed for both of these scores.

The between occasions (stability) coefficient was computed for only the aams 3-6 ast(2) 3-6 group, which was measured before and after student teaching in grades 3-6. The interval between testings was about four months. Since student teaching might be interpreted as a treatment it was necessary to observe changes in mean scores before and after student teaching. The means may be observed in Table 36. In order to examine for possible practice effects, the ast(1) 3-6 group, which was measured only once, was used as a control group. The mean of this group may also be observed in Table 36.

TABLE 36. MEANS AND VARIANCES OF PI SCORE, AND t VALUES FOR THREE GROUPS OF STUDENT TEACHERS, ONE GROUP MEASURED BEFORE AND ONE GROUP BEFORE AND AFTER STUDENT TEACHING.

Groups	n	s^2	$\bar{x}$	t	p
aams 3-6	28	17.83	11.75		
				.559	>.60
ast(2) 3-6	28	12.74	12.00		
				.436	>.60
ast(1) 3-6	25	13.58	11.56		

If student teaching is a relevant treatment or if the second measurement is open to significant practice effects from the first measurement, it is not

apparent in Table 36.

The split-half and stability coefficients computed may be observed in Table 37.

TABLE 37. RELIABILITY COEFFICIENTS BY GROUPS, TYPES, AND SCORES.

Group	Reliability Coefficients			
	PI Score		CR Score	
	Split-half	Stability	Split-half	Stability
ast(2) 3-6	.87	.63	.86	.53
Consolidated town-county teachers	.72	---	.48	---
Consolidated city-county teachers	.76	---	.84	---
All teachers	.84	---	.83	---

The reliability of the consistency criterion cannot be adequately expressed as a reliability coefficient, nor can stability be tested by X^2 since correlated groups would be involved. However, some notion of the stability of this criterion may be obtained by examining how many greater or fewer problems each member of the ast(2) 3-6 group solved above mean chance level after student teaching than before student teaching. These data may be observed in Table 35.

TABLE 38. DIFFERENCES IN NUMBER OF PROBLEMS SOLVED ABOVE CHANCE LEVEL BETWEEN FIRST AND SECOND TESTING OF STUDENT TEACHERS.

Group	Magnitude of Difference -3	-2	-1	0	+1	+2	+3	N
ast(2) 3-6	0	0	9	9	7	2	1	28
				89.5%		7.0%	3.5%	

DISCUSSION

Because the central purpose of the present study was to validate a particular set of problems in teaching arithmetic, the central question for discussion is the extent to which the results provide evidence of the validity of the problems.

With respect to the criterion of differentiating teachers as a group from comparably educated non-teachers, either the continuous PI score or the consistency criterion may be regarded as valid. It should be noted, however, that the PI score is not a powerful discriminator in that the distribution of teacher scores greatly overlaps that of the non-teachers. The failure to obtain better differentiation may be partly attributable to the failure to obtain samples precisely comparable in all respects except professional education and elementary school teaching experience. It may also be partly attributable to the small number of problems used. A third possibility is that the level of skills acquired by the general population of teachers in teaching arithmetic does not actually greatly exceed that of the general college population. With reference to the assumption made earlier, that teachers are specialized, it seems distinctly possible that this assumption is true, but that the degree of specialization, beyond a college degree, is very slight for a sizeable proportion of the elementary teacher population.

With respect to the criterion that problems in teaching arithmetic be sensitive to teaching experience a complex state of affairs obtains. While teachers with 1-3 years experience from preparatory institutions with more

than 5,000 students substantially outperform preparatory teachers at the time of graduation, that the performance of the former group is attributable to teaching experience cannot be demonstrated unequivocally by cross-sectional sampling. It is possible, although not highly probable, that the school systems sampled have selected only the very best students for employment. There are two strong arguments against the latter possibility. First, the variance of teachers with 1-3 years experience is somewhat larger than for preparatory teachers, (Table 12). One would expect a reduction in variance if beginning teachers in the systems sampled were a highly select group. Second, in order to obtain a group of neophyte teachers who score as well as those sampled, it would be necessary to assume that the school systems sampled are in a position to attract only the most select beginning teachers, leaving the less proficient neophytes for the several score of systems with whom they compete. That the two systems sampled compete more effectively for beginning teachers than all other systems seems rather improbable.

A second factor which complicates the effects of teaching experience on performance when cross-sectional methods are used is the changes in the characteristics of teachers which seem to occur in the teacher population as years of experience increase. The sources of teachers according to size of preparatory institution clearly undergoes some change as experience increases, in the samples drawn. This fact alone indicates that it is quite possible that teachers with several years of experience represent a population somewhat different from either preparatory or beginning teachers. That the more experienced teachers are the more select group is nowhere substantiated in the results.

Examination of Tables 10, 12, and 13 suggests the possibility that the reverse may in fact be the case.

The questions with respect to the relationship between criterion performance and the several independent variables tested are first, whether any relationships obtain which are inconsistent with the test rationale; and second, whether useful inferences can be made from these relationships. The positive significant correlations between the criterion and intelligence and the criterion and reading comprehension are in line with expectancy. While reading comprehension accounts for between 4% and 20% of the variance of the criterion, depending on the sample, little of the variance of reading comprehension is independent of the variance of intelligence, with which reading comprehension is closely correlated in the samples drawn.

The positive correlation between problem solving performance and the MTAI indicates at least some commonality between these two measures. Whether this commonality is to be found among teachers as well as student teachers and precisely what it is that accounts for the relationship between the two measures remains to be investigated in subsequent studies.

The relationship between institution size and criterion performance needs rather careful interpretation. In the teacher samples drawn, the small colleges represented are not small colleges with national reputations, but rather seem to serve a clientele from a restricted geographical area. The ability of these schools to compete for outstanding students and faculty and to present a full array of course offerings is no doubt also restricted. Each of these factors might contribute to the somewhat lower criterion performance

which seems to characterize the teachers they prepare. Precisely what factors are involved cannot, of course, be determined from the data thus far collected.

A second aspect of interest in the relationship between institution size and criterion performance is that the graduates of the small colleges represented in the sample seem to stabilize at a considerably lower level of performance than teachers from larger institutions with equal experience. While the performance of preparatory teachers from the smaller institutions at the time of graduation is not known, it would necessarily have to be near the random response mean in order for them to show an increase with teaching experience proportionally as great as that of teachers from the larger institutions. This is to say that unless the graduating student teachers from the smaller institutions perform very, very poorly, there is a distinct possibility of a significant interaction between the size of the graduating institution and teaching experience with respect to criterion performance.

Of the variables which either are not related or are not consistently related to criterion performance some such as recency of arithmetic methods and number of courses in mathematics are probably too crude to show a relationship.

To a certain extent identification of the effects of the grade level at which a teacher teaches and the effects of differences in the location at which teaching is done also suffer from crudeness and confounding. While it would probably not be worthwhile to investigate the former two variables in subsequent studies particularly since more refined indices of arithmetic knowledge

are available, the latter two probably should continue to be investigated: grade level because it is directly relevant to the teachers for whom the problems are appropriate, and location because the finding of equivalence of the teachers between two school systems implies nothing about the equivalence of the teacher in all of the school systems not tested.

The failure of arithmetic methods as a treatment to result in improvement in criterion performance is open to a number of interpretations. One is that the learnings which occur in arithmetic methods result in a change in performance only after interacting with teaching experience. A second is that there was insufficient emphasis on using knowledge to solve problems in the methods course. A third is that the responses used in solving problems in arithmetic are made and learned primarily in the context of the elementary classroom and can be influenced little by formal instruction. Which of these interpretations, if any of them, is the most accurate is a matter that will have to be settled by subsequent research.

The inconsistent relationships between the Study of Values and criterion performance suggest that there is no firm basis for supposing that one particular value orientation or another is significant in criterion performance among student teachers. Whether the same is the case among experienced teachers remains to be determined. At the present time, however, it seems likely that the dominant interests in personality are too broadly defined to be of much use in determining what it is that motivates a teacher to acquire the responses relevant to skill in solving the problems he faces in teaching arithmetic.

The relationships of age and sex to criterion performance seems negligible in the samples drawn, although the fact that there is some decrease (not statistically significant) in performance for older teachers suggests that this variable should be at least observed in subsequent studies.

The split-half reliability of the problems is reasonably high. However the stability coefficient for undergraduates indicates that there is considerable error of measurement between occasions. The latter result suggests that caution should be exercised in using the problems, in their present form, to predict later criterion performance on the basis of criterion performance obtained at an earlier time.

SUMMARY AND CONCLUSIONS

The purpose of the study was to examine the reliability and validity of seven problems in the teaching of arithmetic. The validity of the problems was investigated on the criteria of (1) differentiating elementary school teachers of arithmetic grades 3-6 from comparably educated non-teachers, (2) being sensitive to the effects of elementary school teaching experience, and (3) holding sensible relationships to numerous independent variables. It was found that teachers, as a group, outperform non-teachers, as a group, and that performance is sensitive to the effects of teaching experience. Intelligence, reading comprehension, MTAI score, and size of institution at which a teacher prepares are independent variables positively related to performance. Dominant interests in personality as measured in the Study of Values; location in which teaching is done, an arithmetic methods course as a treatment, recency with which arithmetic methods were taken; number of courses in mathematics; grade level taught; age and sex were found to be unrelated, or not consistently related as determined by cross-validation, to criterion performance.

The split-half reliability for the most valid score was .84 among teachers and .87 among student teachers. The stability coefficient for student teachers, with a 4 month interval between testings, was .63.

It may be concluded from these results that the problems in teaching arithmetic are reasonably reliable, although somewhat weak in stability, and that they hold the expected set of relationships to the criteria chosen for the initial steps in validation.

REFERENCES

1. Allport, G. W.; Vernon, P.E.; and Lindzey, Gardner, Study of Values, Riverside Press, Cambridge, 1960, 3rd Ed.

2. Breuckner, L.J., and Grossnickle, F.E., How to Make Arithmetic Meaningful, John C. Winston & Co., Philadelphia, 1947.

3. Breuckner, L.J.; Merton, E.L.; and Grossnickle, F.E., Discovering Numbers, Exploring Numbers, Understanding Numbers, John C. Winston & Co., Philadelphia, 1952.

4. Brownell, William A., "The Evaluation of Learning in Arithmetic", Arithmetic in General Education, Sixteenth Yearbook of the National Council of Teachers of Mathematics, Bureau of Publications, Teachers College, Columbia University, New York, 1941.

5. Buswell, G.T.; Brownell, W.A.; and Sauble, Irene. A Chart for grades 3-8 of the new arithmetic series Arithmetic We Need, Ginn and Co., Boston, 1954.

6. Buswell, G.T.; Brownell, W.A.; and Sauble, Irene, Arithmetic We Need Books 3-6, Ginn and Co., Boston, 1956.

7. Clark, J.R.; Junge, C.W.; and Moser, H.W., Growth in Arithmetic, Books 3-6, World Book Company, Yonkers-on-Hudson, New York, 1956.

8. Cook, W.W.; Leeds, D.H.; and Callis, Robert, Minnesota Teacher Attitude Inventory, The Psychological Corporation, New York, 1951.

9. Davis, C.C., and Davis T.G., Cooperative English Test, Test C_2, Reading Comprehension Higher Level Form Z, Educational Testing Service, Cooperative Test Division, Princeton, N. J., 1953.

10. Feller, William, An Introduction to Probability Theory and its Applications, Vol. I, John Wiley and Sons, New York, 1950.

11. Hartung, M.L. et. al., Seeing Through Arithmetic, Books 3-4, Scott, Foresman and Co., New York, 1956.

12. Meehl, P.E., and Rosen, Albert, "Antecedent Probability and the Efficiency of Psychological Signs, Patterns or Cutting Scores," Psychological Bulletin, 52: 194-216, May, 1955.

13. Thurstone, L.L., and Thurstone, T.G., American Council on Education Psychological Examination For College Freshman, Educational Testing Service, Cooperative Test Division, Princeton, N. J., 1952.

14. Turner, R.L., and Fattu, N.A., Skill in Teaching, A Reappraisal of the Concepts and Strategies in Teacher Effectiveness Research. Bulletin of the School of Education, Indiana University, Vol. 36, No. 3, May, 1960.

15. Turner, R.L. and Fattu, N.A., Problem Solving Proficiency Among Elementary School Teachers, 1. The Development of Criteria Monograph of the Institute of Educational Research, May, 1960.

16. U.S. Department of Health, Education, and Welfare, Education Directory, 1957-1958, Part 3, Higher Education, United States Government Printing Office, Washington, D.C., 1957.